Assemblée annuelle de la SSC, juin 2003
Recueil de la Section des méthodes d'enquête

AN APPLICATION OF THE BOOTSTRAP VARIANCE ESTIMATION METHOD TO THE PARTICIPATION AND ACTIVITY LIMITATION SURVEY

Dany Faucher, Éric R. Langlet[1] and Éric Lesage[1]

ABSTRACT

The bootstrap method is increasingly used to estimate the variance of estimates obtained from complex survey designs. This method has been shown to work well for a wide range of estimators, including medians and quantiles, as well as smooth functions based on totals. In addition, the bootstrap can be less computer intensive than the jackknife method for surveys with a very large number of primary sampling units (PSUs). The sampling plan of the Canadian 2001 Participation and Activity Limitations Survey (PALS) is a stratified two-stage design in which PSUs are selected without replacement with probability proportional to size. The survey presents specific challenges to the use of the bootstrap method. For instance, the sampling fraction for PALS is relatively high in many strata, which causes the bootstrap method to overestimate the variance. What is the magnitude of this overestimation? Also, a logistic regression response propensity model is used for the nonresponse adjustment in PALS. Should a logistic regression model be fitted to each bootstrap sample? How does this method compare with maintaining fixed response classes over all bootstrap samples? This paper will address these issues.

KEY WORDS: bootstrap, without replacement design, response propensity model, logistic regression.

RÉSUMÉ

La méthode bootstrap est de plus en plus utilisée pour estimer la variance des estimations pour des enquêtes à plan de sondage complexe. Cette méthode possède l'avantage de pouvoir être applicable à toute une gamme de statistiques, incluant les médianes et les quantiles ainsi que les fonctions lisses basées sur les totaux. De plus, l'application du bootstrap est plus rapide en termes de calcul que celle du Jackknife pour des enquêtes avec un grand nombre d'unités primaires d'échantillonnage (UPE). Le plan de sondage de l'Enquête canadienne sur la Participation et les Limitations d'Activités 2001 (EPLA) est un plan stratifié à deux degrés où les UPE sont tirées sans remise avec probabilité proportionnelle à la taille. L'enquête présente des défis intéressants en ce qui a trait à l'utilisation du bootstrap. D'abord, les fractions de sondage pour l'EPLA sont relativement élevées dans certaines strates, ce qui implique une surestimation de la variance avec le bootstrap. Quelle est l'amplitude de cette surestimation? Aussi, un modèle de régression logistique a été utilisé pour l'ajustement de la non-réponse de l'EPLA. Doit-on ajuster un modèle à chacun des échantillons bootstrap ou bien peut-on utiliser des classes de non-réponse fixes pour tous les échantillons bootstrap? Ces questions sont discutées dans cet article.

MOTS CLÉS: bootstrap, tirage sans remise, modèle de propension de la réponse, régression logistique.

1. SAMPLING PLAN OF PALS

The 2001 Participation and Activity Limitation Survey (PALS) collects information about Canadian residents whose everyday activities are limited because of a health-related condition or problem. The survey provides essential information on the prevalence of various disabilities, the supports for people with disabilities, their employment profiles, their income and their participation in society. This information will be used by all levels of government, associations, researchers and non-government organisations to support the planning of services needed by people with activity limitations in order to participate fully in society.

PALS is referred to as a post-censal survey because it uses the Census as a sampling frame to identify its target population. The 2001 Canadian Census long form, which is administered to a one in five sample of households, contains two general filter questions on activity limitations and long-term disabilities. The 2001 PALS selected a sample of individuals from respondents on the Census long-form questionnaire who reported a positive response to at least one of

[1] Dany Faucher and Éric R. Langlet, Statistics Canada, R.H.Coats building, 15Q, Tunney's Pasture, Ottawa, Canada, K1A 0T6, and Éric Lesage, Insee - Direction Interrégionale Antilles-Guyane, 41, rue Bébian, BP 300, 97 158 Point-A-Pitre Cedex, Guadeloupe

these two filter questions. The sampling frame used for PALS consists of estimates of the 2001 Census disabled population within each Enumeration Area (EA), age group, and severity of disability (individuals limited "Often" and individuals limited "Sometimes"). These estimates were obtained from demographic projections of the Canadian population to which disability rates (using the Census definition) estimated from the 2000 PALS pilot test were applied. Census EAs are small geographical data collection units.

The strata are defined by the cross-classification of the ten provinces, four age groups and the Census severity. For the purpose of sample selection, each stratum is subdivided into potentially three sub-strata according to PSU sizes (small, medium and large). Independent samples are geographically selected within each sub-stratum. The PSUs are made up of one or more EAs. The PSU size is predicted from the projected Census disabled population for the combination of EAs, age group and severity corresponding to the PSU. The sample design is a two-stage stratified design that uses the 2001 Census long-form sample in the second stage. PSUs for which the predicted number of disabled individuals was very small or null (small PSUs) were selected by stratified simple random sampling (stratified SRS). Medium PSUs were selected without replacement using probability proportional-to-size (PPS) sampling (in fact, it is really probability proportional to the estimated size sampling, but for simplicity, it will be referred as PPS sampling). Large PSUs were selected with probability one (*take-all* PSUs). *Take-all* PSUs occur for two reasons. First, a *take-all* of PSUs can be required in small strata. Second, the relative size of the PSU within a stratum can be too large to be selected with probability less than one in PPS sampling. In the second stage of the sample design, all Census long-form respondents in a selected PALS PSU are included in the 2001 PALS sample. The total sample size for PALS is about 43,000 individuals.

2. APPLICATION OF THE BOOTSTRAP METHOD FOR PALS

The variance estimation for PALS was done using the bootstrap method of Rao and Wu (1988). In order to simplify the operational aspects of this method, bootstrap weights were used (Rao *et al.*, 1992). For PALS, 500 with replacement samples, called bootstrap samples, were selected from the original sample. For a given bootstrap sample, the initial sampling weight of an individual is adjusted as a function of the bootstrap sample sampling fraction as well as the number of times the individual was selected in the bootstrap sample. These initial bootstrap weights are then adjusted for each weighting step performed in the full sample. This method incorporates the variance component coming from each weight adjustment. The empirical variance of these estimates over all bootstrap samples is used as an estimate of the variance of the estimate. Bootstrap samples were selected in three situations, depending on the first stage sampling fraction and the second stage sampling fraction of households who received the long form.

The first situation corresponds to *take-some* PSUs (small and medium PSUs). The second situation corresponds to *take-all* PSUs but where households at the second stage were selected with probability less than one. The third situation applies to individuals selected with probability one (*take-all* at both stages). A handful of individuals are in this situation, which will be ignored in this paper.

For *take-some* PSUs at the first stage, no distinction for bootstrap sampling is made regarding the second-stage sampling fraction. More variation is expected between totals of PSUs subsampled at the second stage than between totals of PSUs with full enumeration at the second stage. In sub-strata composed of either small or medium PSUs (*take-some* sub-strata), a sample of $n_h - 1$ PSUs within n_h PSUs is selected with replacement for each bootstrap sample. The particular selection of $n_h - 1$ PSUs within n_h PSUs simplifies the bootstrap weight formula (Rao, Wu, Yue, 1992). All full-sample second-stage units of the $n_h - 1$ selected PSUs are in the bootstrap sample.

Let w_{hij} denote the initial sampling weight of the j^{th} individual in the i^{th} PSU of the h^{th} sub-stratum. For a given bootstrap sample, the initial bootstrap weights are given by

$$w_{hij}^{B} = \frac{n_h}{n_h - 1} m_{hi}^{*} w_{hij} \tag{1}$$

where m_{hi}^{*} represents the number of times that the hi^{th} PSU is selected in the bootstrap sample. These adjustments to the w_{hij} are used so that the sum of the bootstrap weights estimates the population total.

For sub-strata composed of *take-all* PSUs (large PSUs) at the first stage, the sample design (a stratified systematic sample of households) can be approximated by a single-stage stratified SRS of individuals. In each such sub-stratum *h*, a with-replacement random sample of $m_h - 1$ individuals within m_h individuals is selected for each bootstrap sample. The initial bootstrap weights are given by

$$w_{hij}^{B_I} = \frac{m_h}{m_h - 1} m_{hij}^{*} w_{hij} \tag{2}$$

where m_{hij}^{*} represents the number of times that the hij^{th} individual is selected in the bootstrap sample. In this formula, the PSU subscript is only used to classify the individuals within their original PSU.

Initial bootstrap weights have to be adjusted in the same way as the initial weights of the original sample. The next step was the nonresponse adjustment which is described in more detail in Section 4. The last step of the weight adjustment was the post-stratification to Census totals estimated from the roughly one-in-five systematic sample of households that received the long form. The weights of each bootstrap sample were post-stratified so that the sum of the weights for each post-stratum would add up to the estimate from the Census.

This bootstrap method is applicable to sample designs in which PSUs are selected with replacement or cases where the PSU sampling fraction is small in most strata. Omission to these conditions could lead to an overestimation. A comparative study was done to measure the extent of this overestimation.

3. COMPARISON OF THE YATES-GRUNDY AND BOOTSTRAP ESTIMATORS

A comparison was done between the bootstrap estimator and a theoretical estimator of the variance for the PALS sample design. For this comparison some simplifying assumptions about the sample design were made. The first stage of the PALS sample design is a stratified SRS of small PSUs, a stratified PPS sample of medium PSUs and a *take-all* sample of large PSUs. At the second stage, it will be assumed that the PALS sample design can be approximated by a stratified SRS of disabled individuals. It is in fact a stratified systematic sample of households. This last assumption is reasonable if the disabled population is distributed uniformly throughout each PSU. The comparison was done using only the sampling weights and therefore, do not take into account nonresponse and post-stratification adjustments.

For all PSUs of a given sub-stratum, the variance estimate for an SRS of m_i within M_i disabled individuals within PSU *i* is given by:

$$v(\hat{Y}) = \sum_{i \in s_i} \frac{1}{\pi_i} M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i} \right) s_i^2 \tag{3}$$

where π_i is the inclusion probability of the i^{th} PSU and s_i^2 is the within PSU sample variance. A between PSU variance estimator for any sample of PSUs under any fixed-size design at the first stage can be obtained from the Yates-Grundy (1953) estimator. Combining this estimator with expression (3) gives, for a particular sub-stratum, an estimate of the variance for an estimated total $\hat{Y}$ (Särndall *et al.*, 1992):

$$v_1(\hat{Y}) = \frac{1}{2} \sum_{i \neq j}\sum \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{\hat{Y}_i}{\pi_i} - \frac{\hat{Y}_j}{\pi_j} \right)^2 + \sum_{i \in s} \frac{1}{\pi_i} M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i} \right) s_i^2 \tag{4}$$

where π_{ij} are joint inclusion probabilities for the i^{th} and j^{th} PSU. In sub-strata composed of small PSUs the left-hand side of expression (4) reduces to the variance of a simple random sample of PSUs. In sub-strata composed of medium PSUs, PPS sampling was done through the use of PROC SURVEYSELECT of SAS with the procedure of Hanurav (1967) and Vijayan (1968). In sub-strata composed of *take-all* PSUs, the left-hand side of expression (4) becomes zero because the sample design for these sub-strata reduces to a single-stage design with units selected by stratified simple random sampling. For the purpose of this comparison, some simplifications were made. For the purpose of this comparison, some simplifications were made. PSUs composed only of survey non-respondents or PSUs including only respondents with no limitation (in both cases $\hat{Y}_i = 0$) were assigned a zero within PSU variance (right-hand side of expression (4)). Obviously, PSUs with disabled individuals but for which no one has the characteristic of interest have also a null within PSU variance. Since the size M_i of each PSU is unknown, an estimate $\hat{M}_i$ was used from the long-form Census sample.

For linear statistics such as totals, means, et cetera, the bootstrap variance formula can be approximated by the variance formula of a PPS with-replacement sample of PSU at the first stage with any design at the second stage.

$$v_2(\hat{Y}) = \frac{1}{n(n-1)} \sum_{i \in s} \left(\frac{\hat{Y}_i}{p_i} - \hat{Y} \right)^2 \qquad (5)$$

where p_i is the probability of selection of the i^{th} PSU. See, for instance, Cochran (1977) for expression (5). For linear statistics, both variance estimate formula are unbiased for the true variance $V(\hat{Y})$ but are not equivalent. The precision of the bootstrap estimator approaches the one of $v_2(\hat{Y})$ as the number of bootstrap samples tends to infinity (Rust and Rao, 1996). Since both estimators are unbiased, the bootstrap variance estimate will also approach $v_2(\hat{Y})$ for an infinite number of bootstrap samples. Since our comparison was restricted to linear statistics, it was not necessary to generate the bootstrap samples. Instead, the bootstrap variance formula in a particular sub-stratum of small or medium PSUs was approximated by the expression in (5).

For small PSUs (drawn by SRS), $p_i = 1/N$ and the formula reduces to the variance of a simple with-replacement random sample of PSUs. For medium PSUs, p_i is the relative size of the PSU within the sub-stratum. For large PSUs (take-all), the formula is replaced by the variance of a simple with-replacement random sample of m from M disabled individuals.

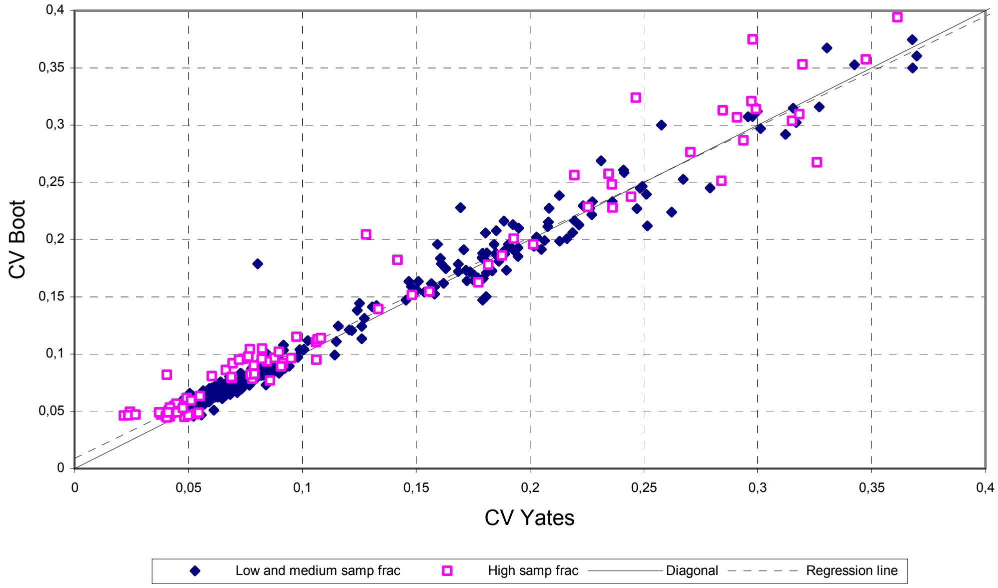
Variance estimates $v_1(\hat{Y})$ and $v_2(\hat{Y})$ were calculated for several PALS variables, such as being disabled, using of hearing aids, et cetera. Estimation was done by province and age group. Since the prevalence of these limitations differs substantially from one limitation to another, variance estimates were compared in terms of coefficients of variation (CV). A regression analysis was performed to summarize the results. In order to evaluate the extent of the overestimation for strata with large PSU sampling fractions (the term "strata" here refers to the cross-classification of province, age group and severity), a first regression was done for these specific strata and a second regression was done on the remaining strata. Large sampling fractions were defined as sampling fractions larger than 20%, which constitutes 28% of the strata. A third regression was done for all strata combined. Each unit in the regression corresponds to a particular stratum and a particular estimate. Results are shown on Table 1 below.

The relationship between the CVs obtained from the bootstrap and Yates-Grundy estimators is quite strong for the three analyses, with R^2 values above 96%. The slope is significantly different from one in each case and the intercept is also significantly different from zero in each case (not shown in the Table 1). Although the overestimation seems to be higher for strata with large sampling fractions, particularly for the lower CVs, the differences are not large globally. It seems that there is no systematic bias for $v_2(\hat{Y})$ compared to $v_1(\hat{Y})$, since the intercept is greater than zero but the slope is lower than one. Under the fitted models, the positive intercept combined with the slope slightly smaller than 1 indicates that $v_2(\hat{Y})$ is larger than $v_1(\hat{Y})$ especially for small to moderate CVs. For large CVs, $v_2(\hat{Y})$ is on average slightly smaller than $v_1(\hat{Y})$. This could be explained by the large variability of the estimates $v_1(\hat{Y})$ and $v_2(\hat{Y})$ for large CVs corresponding to rare characteristics or could be partially due to the simplifying assumptions that were made. It was also found as expected that, on average, the higher the sampling fraction at the PSU level the larger the size of the over-estimation for the bootstrap method. It should be noted, however, that the bootstrap estimator was compared to the Yates Grundy estimator, which is not the true variance. It is rather a "reasonable" direct estimate of the true variance which does not assume that the PSUs are drawn with replacement. Therefore, the term "overestimation" should be used with caution since two estimates are compared. Figure 1 presents the fitted regression on all strata.

Table 1. Results of the regression $CV_{boot} = \beta_0 + \beta_1 CV_{yates}$

Model	β_0	β_1	R^2
Strata with large sampling fraction (>20%)	0.0157	0.9502	0.9669
Other strata	0.0065	0.9719	0.9770
All strata	0.0089	0.9657	0.9740

Figure 1. Regression on all the strata



4. ADJUSTMENT OF BOOTSTRAP WEIGHTS FOR NONRESPONSE

The full-sample weights were adjusted for nonresponse through the use of a response propensity model, a logistic regression model that predicts the response probability. The method used is the method of equal deciles where individuals (respondents and non-respondents) are grouped into 10 classes of approximately equal sizes, homogeneous with respect to their predicted probability of response. A criterion used in the logistic regression was to minimize the Hosmer and Lemeshow statistic (Hosmer and Lemeshow, 1989).

This step involves quite a bit of modelling because it requires the fitting of many different models to find the most appropriate model. Since this aspect requires manual intervention, it is not practically possible to completely remodel each bootstrap sample. On the other hand, since all bootstrap samples come from the same main sample, they should follow the model used for the main sample. The second-best option would be to re-estimate the same model on each bootstrap sample (Method A). New parameter estimates would be produced for each bootstrap sample, leading to new nonresponse classes and new adjustment factors. Since the fitting of a logistic regression model is an iterative process, this method still can be quite computer intensive when conducted on 500 different samples. Moreover, there is no guarantee that convergence will be attained for all bootstrap samples.

An alternative to this method would be to maintain fixed nonresponse classes over all bootstrap samples (Method B). The bootstrap samples being different from one another, the number of respondents and non respondents within each class would be different and different nonresponse adjustment factors would be applied to each bootstrap sample.

A comparative study was done to measure the differences between Method A and B. The CVs of certain key statistics are compared in Table 2. This table presents estimates of children aged 5 to 14 with vision disability. The table is broken down by severity level. The CVs for both methods and the relative differences are presented.

Table 2. Estimates and CVs of children aged 5-14 with vision disability

Severity	Estimate	CV (A)	CV (B)	Rel. diff. (%)
Mild	1,412	20.575	20.482	0.45
Moderate	4,168	15.567	15.614	-0.30
Severe	3,922	13.660	13.645	0.11
Very severe	5,012	13.905	13.867	0.27

As can be seen from this study, it is not necessary to perform the weight adjustments in the bootstrap sample in exactly the same way as this was done for the full sample. In this example, the relative difference between the CVs obtained from both methods is clearly negligible, the maximum difference being less than 0.5%. Other tables not included in this paper showed comparable relative differences.

5. SUMMARY AND CONCLUSION

Specific challenges to the application of the bootstrap variance estimation method to PALS were presented in this paper. In particular, the relatively high PSU sampling fraction in some of the strata could cause the bootstrap method to overestimate the variance. A study was done to compare the bootstrap estimate to an approximate direct variance estimate using the Yates-Grundy estimator. A few simplifications had to be done to make this comparison. Results indicated that although the bootstrap variance estimator tends to overestimate the variance, the relative size of this over-estimation is relatively small. Therefore, the bootstrap method was judged acceptable for PALS. In this case, it was felt that the variance estimates were in general only slightly conservative, which is preferable to the opposite. Of course, these results are limited to our particular application.

Another challenge in using the bootstrap method was the fact that a response propensity model was used for the nonresponse adjustment of the main sample. Re-estimating the parameters of the same logistic regression model on each bootstrap sample can be quite computer intensive and convergence may not be obtained on all bootstrap samples. This may require simplifying the initial logistic regression model to accommodate all bootstrap samples. The drawback is a loss of optimality for the nonresponse adjustment of the full sample. The study showed that instead of re-estimating the same logistic regression model on each bootstrap sample, an alternative approach where the nonresponse classes are fixed over all bootstrap samples gives almost exactly the same results in terms of variance estimation. Therefore, the alternative approach is recommended to avoid excessive computer time and possible convergence problems.

REFERENCES

Cochran, W.G. (1977). *Sampling Techniques*, New York: John Wiley & Sons, 254.

Hanurav, T.V. (1967). *Introduction to Survey Sampling*, Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 01-035, Beverly Hills and London: Sage Publications Inc.

Hosmer, D.W. Jr. and S. Lemeshow (1989). *Applied Logistic Regression*, New York: John Wiley & Sons.

Rao, J.N.K. and Wu, C.F.J. (1988). Resampling Inference With Complex Survey Data. *Journal of the American Statistical Association*, **83**, 231-241.

Rao, J.N.K., Wu, C.F.J. and Yue, K. (1992). Some Recent Work on Resampling Methods for Complex Surveys. *Survey Methodology*, **18,** 209-217.

Rust, K.F. and Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, **5,** 283-310.

Sarndäl, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag New-York.

SAS Institute Inc. (1999). *SAS/STAT User's Guide, Version 8*, Vol. 2, p. 3273-3320.

Vijayan, K. (1968). An exact PPS Sampling Scheme: Generalization of a method of Hanurav. *Journal of the Royal Statistical Society*, Series B, **30**, 556-566.

Yates, F. and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, Series B, **15**: 235-261.